

Computer Science Department

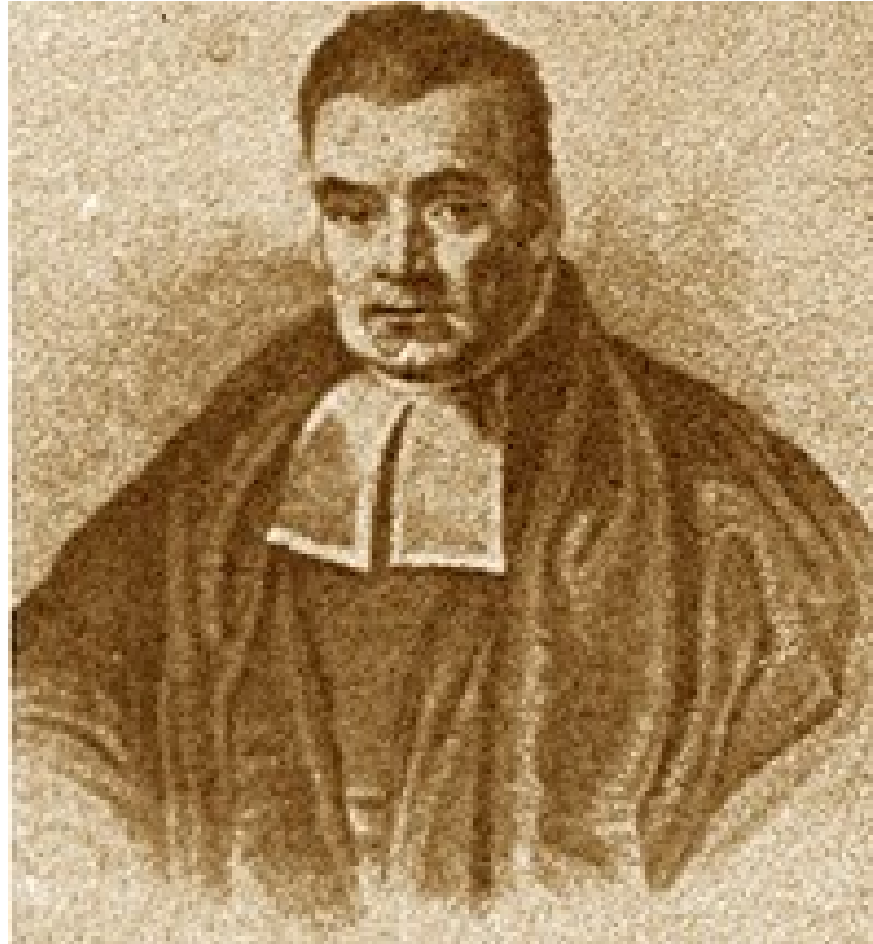
University of Verona

A.A. 2017-18

Pattern Recognition

Bayes decision theory

Rev. Thomas Bayes, F.R.S (1702-1761)



Introduzione

- Approccio statistico fondamentale di classificazione di pattern
- Ipotesi:
 1. Il problema di decisione è posto in termini probabilistici;
 2. Tutte le probabilità rilevanti sono conosciute;
- Goal:

Discriminare le differenti *regole di decisione* usando le *probabilità* ed i *costi* ad esse associati;

Un esempio semplice

- Sia ω lo *stato di natura* da descrivere probabilisticamente;
- Siano date:
 1. Due classi ω_1 and ω_2 per cui sono note
 - a) $P(\omega = \omega_1) = 0.7$
 - b) $P(\omega = \omega_2) = 0.3$

= **Probabilità a priori** o **Prior**
 2. Nessuna misurazione.
- Regola di decisione:
 - Decidi ω_1 se $P(\omega_1) > P(\omega_2)$; altrimenti decidi ω_2
- Più che decidere, *indovino* lo stato di natura.

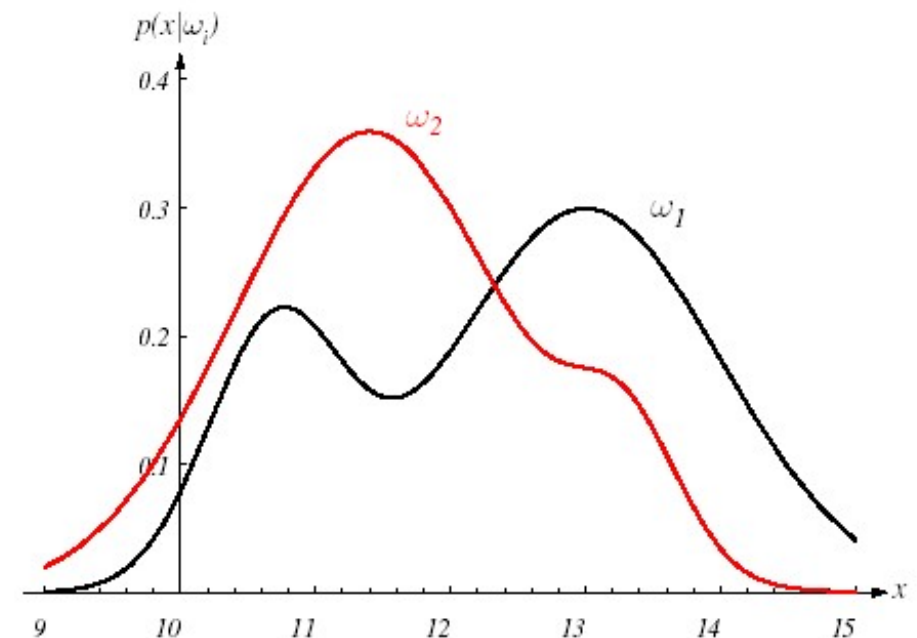
Altro esempio – Formula di Bayes

- Nell'ipotesi precedente, con in più la singola misurazione x , v.a. dipendente da ω_j , posso ottenere

$$p(x | \omega_j)_{j=1,2} = \text{Likelihood, o Probabilità stato-condizionale}$$

ossia *la probabilità di avere la misurazione x sapendo che lo stato di natura è ω_j*

Fissata la misurazione x più è alta $p(x | \omega_j)$ più è probabile che ω_j sia lo stato “giusto”.



Altro esempio – Formula di Bayes (2)

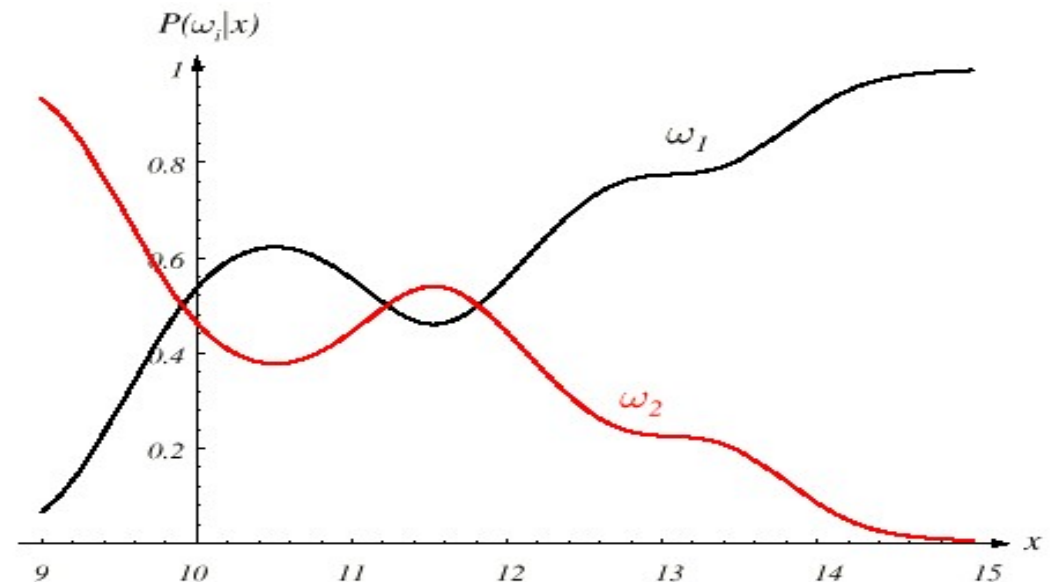
- Note $P(\omega_j)$ e $p(x | \omega_j)$, la decisione dello stato di natura diventa, per Bayes

$$p(\omega_j, x) = P(\omega_j | x)p(x) = p(x | \omega_j)P(\omega_j)$$

ossia

$$P(\omega_j | x) = \frac{p(x | \omega_j)P(\omega_j)}{p(x)} \propto p(x | \omega_j)P(\omega_j), \text{ dove:}$$

- $P(\omega_j)$ = Prior
- $P(x | \omega_j)$ = Likelihood
- $P(\omega_j | x) = \textit{Posterior}$
- $p(x) = \sum_{j=1}^J p(x | \omega_j)P(\omega_j)$
= *Evidenza*



Regola di decisione di Bayes

$$P(\omega_j | x) = \frac{p(x | \omega_j)P(\omega_j)}{p(x)} \iff \text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

- Ossia il *Posterior* o **probabilità a posteriori** è la probabilità che lo stato di natura sia ω_j data l'osservazione x .
- Il fattore più importante è il prodotto *likelihood* \times *prior* ;
l'evidenza $p(x)$ è semplicemente un fattore di scala, che assicura che

$$\sum_j P(\omega_j | x) = 1$$

- Dalla formula di Bayes deriva **la regola di decisione di Bayes:**

Decidi ω_1 se $P(\omega_1|x) > P(\omega_2|x)$, ω_2 altrimenti

Regola di decisione di Bayes (2)

- Per dimostrare l'efficacia della regola di decisione di Bayes:
 - 1) Definisco la *probabilità d'errore* annessa a tale decisione:

$$P(error | x) = \begin{cases} P(\omega_1 | x) & \text{se decido } \omega_2 \\ P(\omega_2 | x) & \text{se decido } \omega_1 \end{cases}$$

- 2) Dimostro che ***la regola di decisione di Bayes minimizza la probabilità d'errore.***

Decido ω_1 se $P(\omega_1 | x) > P(\omega_2 | x)$ e viceversa.

- 3) Quindi se voglio ***minimizzare la probabilità media di errore*** su tutte le osservazioni possibili,

$$P(error) = \int_{-\infty}^{+\infty} P(error, x) dx = \int_{-\infty}^{+\infty} P(error | x) p(x) dx$$

se per ogni x prendo $P(error|x)$ più piccola possibile mi assicuro la probabilità d'errore minore (come detto il fattore $p(x)$ è influente).

Regola di decisione di Bayes (3)

In questo caso tale probabilità d'errore diventa

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)];$$

Questo mi assicura che la regola di decisione di Bayes

*Decidi ω_1 se $P(\omega_1|x) > P(\omega_2|x)$, ω_2 altrimenti
minimizza l'errore!*

- ***Regola di decisione equivalente:***

- La forma della regola di decisione evidenzia *l'importanza della probabilità a posteriori*, e sottolinea *l'ininfluenza dell'evidenza*, un fattore di scala che mostra quanto frequentemente si osserva un pattern x ; eliminandola, si ottiene la equivalente regola di decisione:

Decidi ω_1 se $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$, ω_2 altrimenti

Teoria della decisione

- Il problema può essere scisso in una fase di *inferenza* in cui si usano i dati per addestrare un modello $p(\omega_i|\mathbf{x})$ e una seguente fase di *decisione*, in cui si usa la posterior per fare la scelta della classe
- Un'alternativa è quella di risolvere i 2 problemi contemporaneamente e addestrare una funzione che mappi l'input \mathbf{x} direttamente nello spazio delle decisioni, cioè delle classi \rightarrow *linear machine*, che usa *funzioni discriminanti lineari*
- Poniamoci in un caso di classificazione multiclasse, con C classi

Funzioni discriminanti

Esempio per C classi:

- Uno dei vari metodi per rappresentare classificatori di pattern consiste in un set di *funzioni discriminanti* $g_i(\mathbf{x})$, $i=1 \dots C$
- Il classificatore finale, ossia la **linear machine** assegna il vettore di feature \mathbf{x} alla classe ω_i se
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \text{ per ogni } j \neq i$$
- Di per sé l'applicazione della regola di Bayes non permette di osservare il confine di separazione
- In pratica, una linear machine mi permette di visualizzare il confine di decisione in maniera analitica, grazie alle funzioni discriminanti, date alcune assunzioni sulla forma della likelihood e dei prior, che vedremo

Funzione discriminanti (2)

- Esistono molte funzioni discriminanti equivalenti. Per esempio, tutte quelle per cui i risultati di classificazione sono gli stessi
 - Per esempio, se f è una funzione monotona crescente, allora

$$g_i(\mathbf{x}) \Leftrightarrow f(g_i(\mathbf{x}))$$

- Alcune forme di funzioni discriminanti sono più semplici da capire o da calcolare

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_{j=1}^c p(\mathbf{x}|\omega_j)P(\omega_j)}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i)$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i),$$

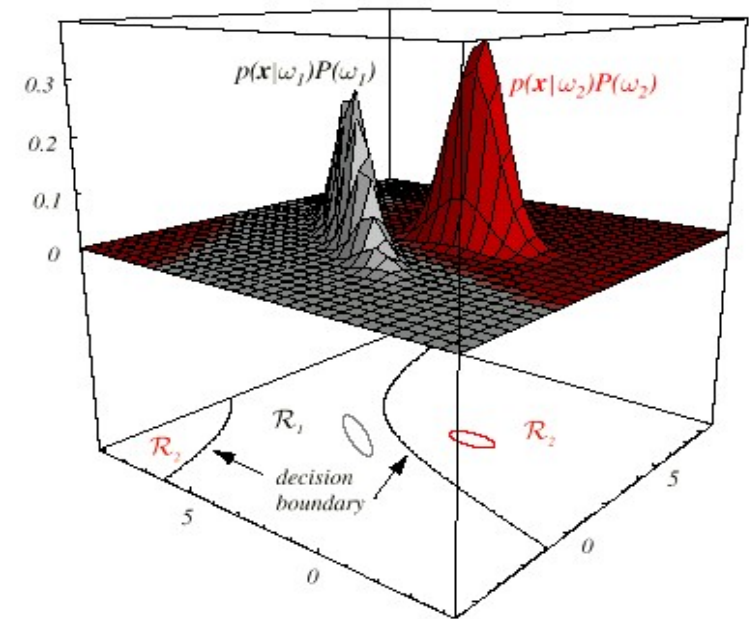
Funzione discriminanti (3)

- L'effetto di una funzione discriminante è quello di *dividere lo spazio delle features* in *c superfici di separazione o decisione*, R_1, \dots, R_c
 - Le regioni sono separate con *confini di decisione*, linee descritte dalle funzioni discriminanti.
 - Nel caso a *due* categorie ho due funzioni discriminanti, g_1, g_2 , per cui assegno x a ω_1 se $g_1(x) > g_2(x)$ o se $g_1(x) - g_2(x) > 0$
 - Quindi

$$g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$$

$$g(\mathbf{x}) = P(\omega_1 | \mathbf{x}) - P(\omega_2 | \mathbf{x})$$

$$g(\mathbf{x}) = \ln \frac{p(\mathbf{x} | \omega_1)}{p(\mathbf{x} | \omega_2)} + \ln \frac{P(\omega_1)}{P(\omega_2)} \quad \text{ottengo una } \textit{linear machine}$$



La densità normale (qui utile per le linear machine... ma anche altrove, come vedremo)

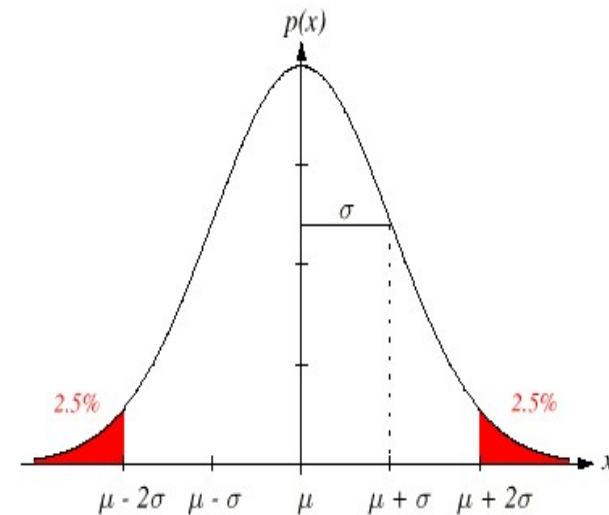
- La struttura di un classificatore di Bayes è determinata da:

- Le densità condizionali $p(\mathbf{x} | \omega_i)$
- Le probabilità a priori $P(\omega_i)$

- Una delle più importanti densità è **la densità normale** o **Gaussiana multivariata**; infatti:

- è analiticamente trattabile;
- fornisce una robusta modellazione di problemi sia teorici che pratici

- il **teorema del Limite Centrale** asserisce che *“sotto varie condizioni, la distribuzione della somma di d variabili aleatorie indipendenti tende ad un limite particolare conosciuto come distribuzione normale”*.



| Intervallo | Inform. |
|---------------------|---------|
| $\mu \pm \sigma$ | 68% |
| $\mu \pm 2\sigma$ | 95% |
| $\mu \pm 2.5\sigma$ | 99% |

Probabilità che il dato sia contenuto negli intervalli di riferimento.

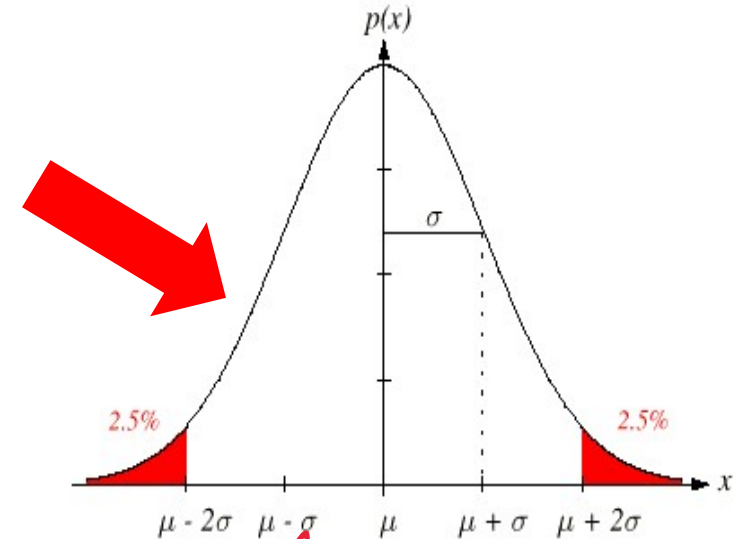
La densità normale (2)

- La funzione Gaussiana ha altre proprietà
 - La trasformata di Fourier di una funzione Gaussiana è una funzione Gaussiana;
 - La moltiplicazione di due funzioni Gaussiane è ancora Gaussiana
 - È ottimale per la localizzazione nel tempo o in frequenza
- Guardate le “Gaussian Identities” o il “Matrix Cookbook”

Densità normale univariata

- Iniziamo con la densità normale univariata. Essa è completamente specificata da due parametri, *media* μ e *varianza* σ^2 , si indica con $N(\mu, \sigma^2)$ e si presenta nella forma

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left\{ -\frac{1}{2} \left(\frac{x - \mu}{\sigma} \right)^2 \right\}$$



Media $\mu = E[x] = \int_{-\infty}^{\infty} xp(x)dx = \frac{1}{N} \sum_{i=1}^N x_i$

Varianza $\sigma^2 = E[(x - \mu)^2] = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx = \frac{1}{N-1} \sum_{i=1}^N (x_i - \mu)^2$

- Fissata media e varianza la densità Normale è quella dotata di massima entropia;
 - L'entropia misura l'incertezza di una distribuzione o la quantità d'informazione necessaria in media per descrivere la variabile aleatoria associata, ed è data da

$$H(p(x)) = -\int p(x) \ln p(x) dx$$

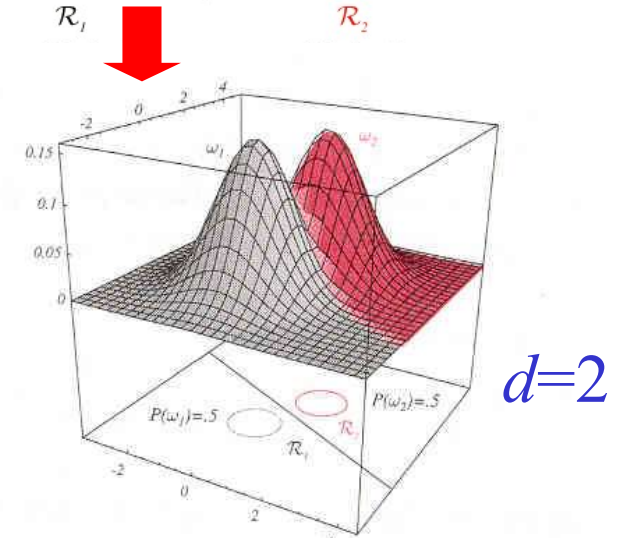
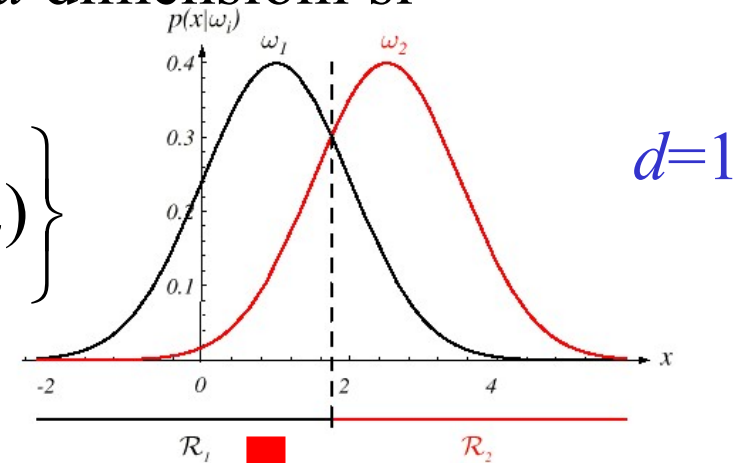
Densità normale multivariata

- La generica densità normale multivariata a d dimensioni si presenta nella forma

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu}) \right\}$$

in cui

- $\boldsymbol{\mu}$ = vettore di **media** a d componenti
- Σ = matrice $d \times d$ di **covarianza**, dove
 - $|\Sigma|$ = determinante della matrice
 - Σ^{-1} = matrice inversa

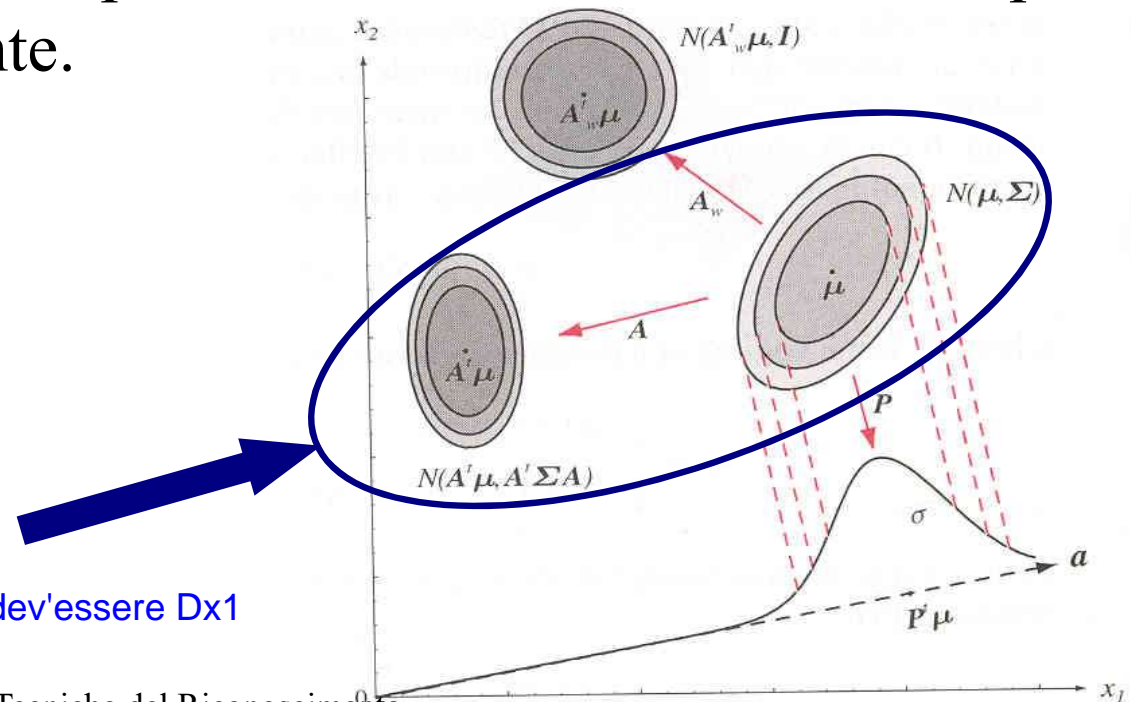


- Analiticamente $\Sigma = E[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t] = \int (\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^t p(\mathbf{x}) d\mathbf{x}$
- Elemento per elemento $\sigma_{ij} = E[(x_i - \mu_i)(x_j - \mu_j)]$

Densità normale multivariata (2)

- Caratteristiche della **matrice di covarianza**
 - Simmetrica
 - Semidefinita positiva ($|\Sigma| \geq 0$) Determinante maggiore o uguale di zero, tutti gli autovalori sono non negativi.
 - σ_{ii} = varianza di x_i ($= \sigma_i^2$)
 - σ_{ij} = covarianza tra x_i e x_j (se x_i e x_j sono *statisticamente indipendenti* $\sigma_{ij} = 0$)
 - Se $\sigma_{ij} = 0 \quad \forall i \neq j$ $p(\mathbf{x})$ è il prodotto della densità univariata per \mathbf{x} componente per componente.
 - Se
 - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \Sigma)$
 - A matrice $d \times k$
 - **$\mathbf{y} = \mathbf{A}^t \mathbf{x}$**
- **$p(\mathbf{y}) \approx N(\mathbf{A}^t \boldsymbol{\mu}, \mathbf{A}^t \Sigma \mathbf{A})$**

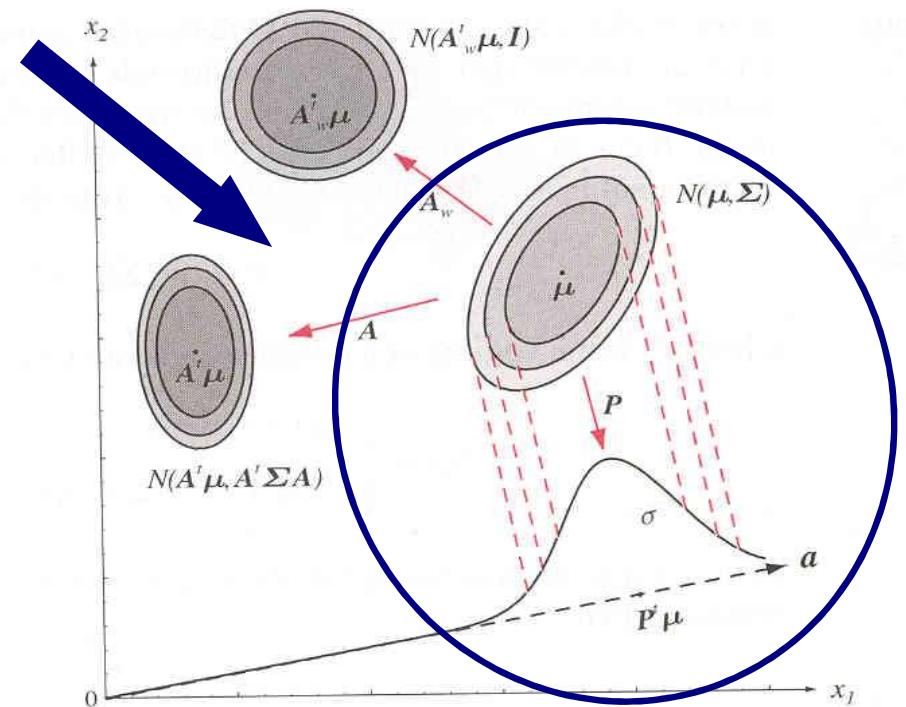
Perché P sia una proiezione unidimensionale, A dev'essere $D \times 1$ così da ottenere un vettore.



Densità normale multivariata (3)

- CASO PARTICOLARE: $k = 1$
 - $p(\mathbf{x}) \approx N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$
 - \mathbf{a} vettore $d \times 1$ di lunghezza unitaria
 - $y = \mathbf{a}^t \mathbf{x}$
 - y è uno scalare che rappresenta la proiezione di \mathbf{x} su una linea in direzione definita da \mathbf{a}
 - $\mathbf{a}^t \boldsymbol{\Sigma} \mathbf{a}$ è la *varianza* di \mathbf{x} su \mathbf{a}

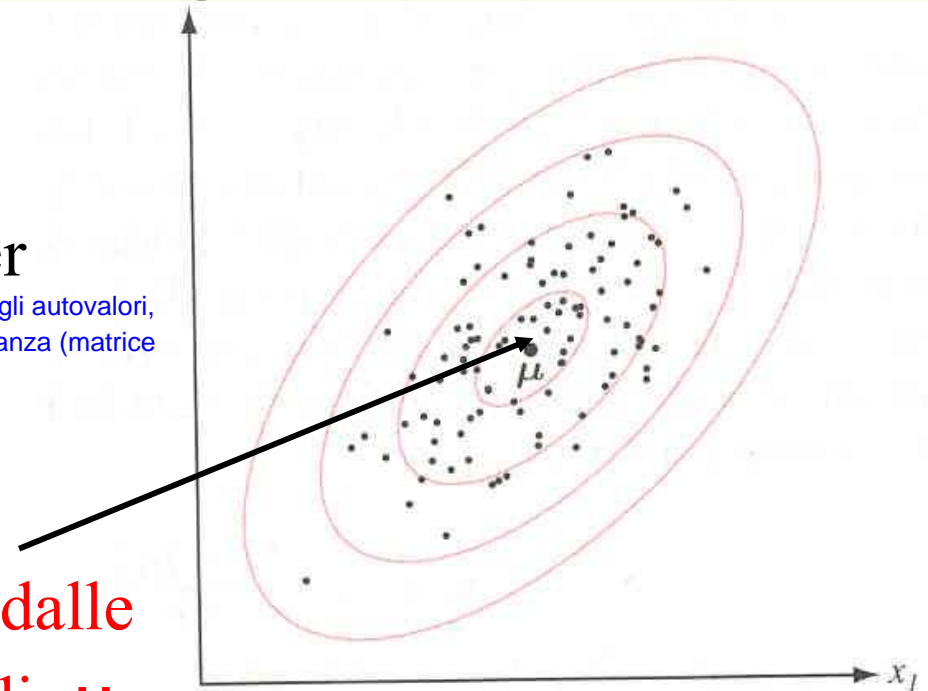
- In generale $\boldsymbol{\Sigma}$ ci permette di calcolare la *dispersione* dei dati in ogni superficie, o sottospazio.



Densità normale multivariata (4) – whitening

- Siano
 - Φ la matrice degli autovettori di Σ in colonna;
 - Λ la matrice diagonale dei corrispondenti autovalori;
- La trasformazione $A_w = \Phi \Lambda^{-1/2}$, applicata alle coordinate dello spazio delle feature, assicura una distribuzione con matrice di covarianza $= I$ (matrice identica)
- La densità $N(\mu, \Sigma)$ d-dimensionale necessita di $d + d(d+1)/2$ parametri per essere definita
 d elementi sono quelli sulla diagonale che contengono gli autovalori, $d(d+1)/2$ sono gli elementi che rappresentano la covarianza (matrice simmetrica, c.a metà elementi)
- Ma cosa rappresentano graficamente Φ e Λ ?

Media
individuata dalle
coordinate di μ



Densità normale multivariata (5)

PCA!!!!!!

Gli assi principali degli iperellissoidi sono dati dagli autovettori di Σ (descritti da Φ)

Gli iperellissoidi sono quei luoghi dei punti per i quali la distanza di \mathbf{x} da μ

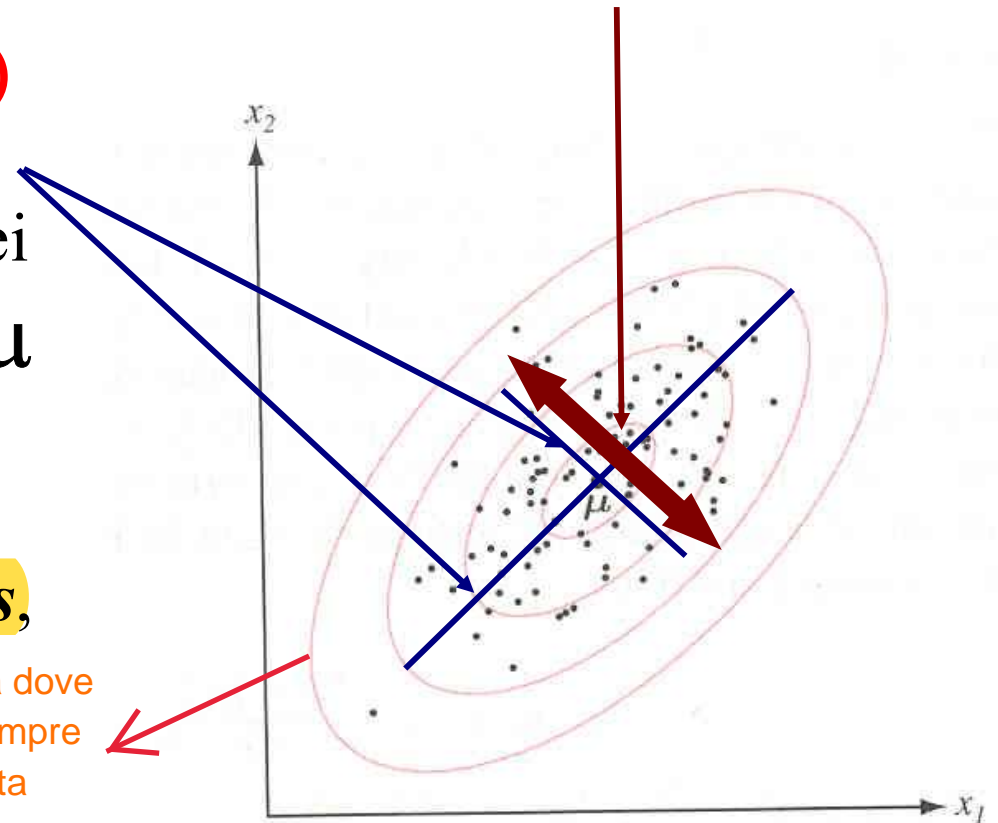
$$r^2 = (\mathbf{x} - \mu)^t \Sigma^{-1} (\mathbf{x} - \mu)$$

detta anche **distanza di Mahalanobis**,
è costante

$$\left(\frac{x - \mu}{\sigma} \right)^2 = 1 - D$$

Punti in cui, non importa dove sono, la probabilità è sempre la stessa. Viene chiamata distanza di Mahalanobis.

Le lunghezze degli assi principali degli iperellissoidi sono dati dagli autovalori di Σ (descritti da Λ)





Funzioni discriminanti - Densità Normale

- Tornando ai classificatori Bayesiani, ed in particolare alle linear machine, analizziamo la funzione discriminante come si traduce nel caso di densità Normale

$$g_i(\mathbf{x}) = \ln p(\mathbf{x} | \omega_i) + \ln P(\omega_i)$$

Riscriviamo la probabilità condizionata in funzione a quanto visto per la distribuzione normale


$$g_i(\mathbf{x}) = \ln \left(\frac{1}{(2\pi)^{d/2} |\Sigma_i|^{1/2}} \exp \left\{ -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right\} \right) + \ln P(\omega_i)$$


$$g_i(\mathbf{x}) = -\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

- A seconda della natura di Σ , la funzione discriminante può essere semplificata. Vediamo alcuni esempi.



Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$

- È il caso più semplice in cui le feature sono statisticamente indipendenti ($\sigma_{ij} = 0, i \neq j$), ed ogni classe ha la stessa varianza (*caso 1-D*):

$$g_i(\mathbf{x}) = -\frac{\|\mathbf{x} - \boldsymbol{\mu}_i\|^2}{2\sigma^2} + \ln P(\omega_i)$$

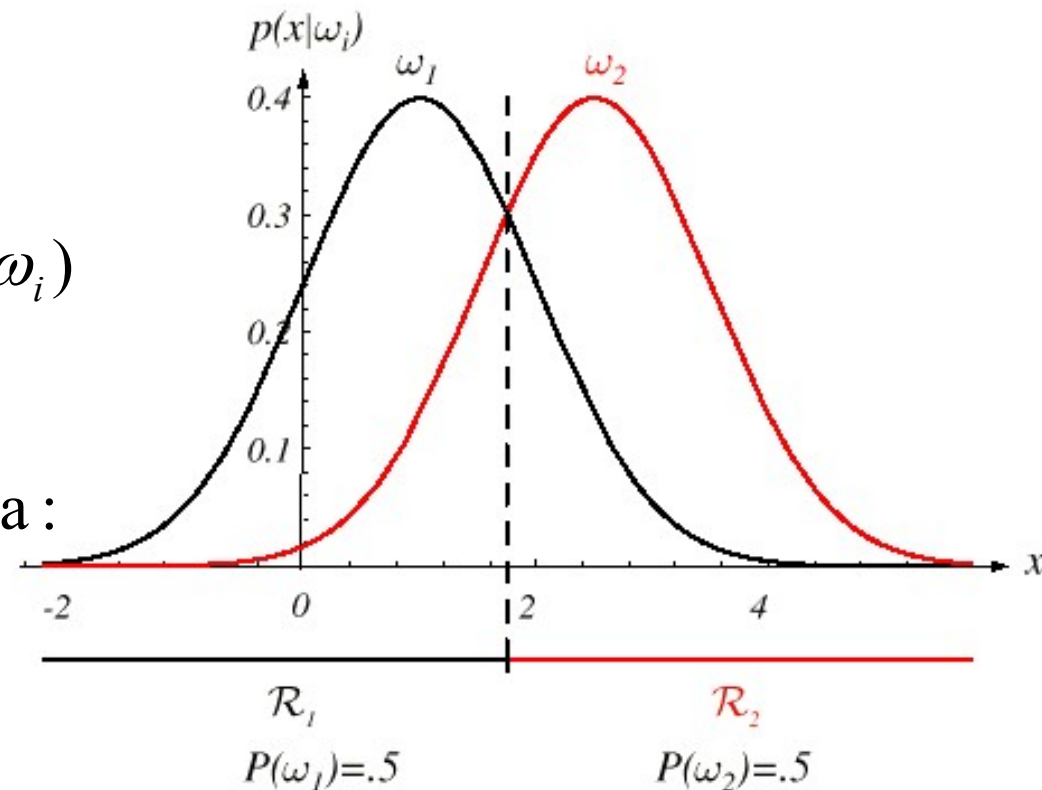
$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2} [\mathbf{x}^t \mathbf{x} - 2\boldsymbol{\mu}_i^t \mathbf{x} + \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i] + \ln P(\omega_i)$$

dove il termine $\mathbf{x}^t \mathbf{x}$, uguale per ogni \mathbf{x} , può essere ignorato giungendo alla forma :

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

dove

$$\mathbf{w}_i = \frac{1}{\sigma^2} \boldsymbol{\mu}_i \quad \text{e} \quad w_{i0} = -\frac{1}{2\sigma^2} \boldsymbol{\mu}_i^t \boldsymbol{\mu}_i + \ln P(\omega_i)$$



Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (2)

- Le funzioni precedenti vengono chiamate *funzioni discriminanti lineari*
- I **confini di decisione** sono dati da $g_i(\mathbf{x}) = g_j(\mathbf{x})$ per le due classi con più alta probabilità a posteriori

– Ponendo $g_i(\mathbf{x}) - g_j(\mathbf{x}) = 0$ abbiamo:


$$\mathbf{w}^t (\mathbf{x} - \mathbf{x}_0) = 0$$

dove

$$\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$$

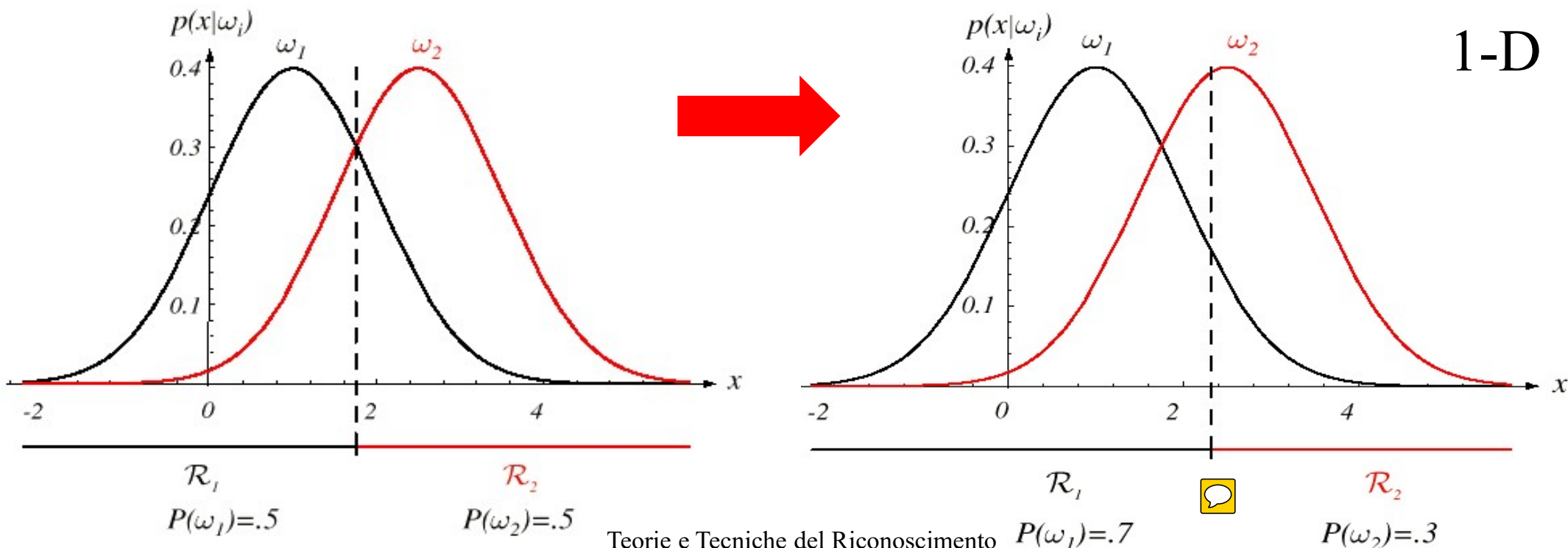
$$\mathbf{x}_0 = \frac{1}{2} (\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

NB: se $\sigma^2 \ll \|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2$
la posizione del confine di
decisione è insensibile ai prior!

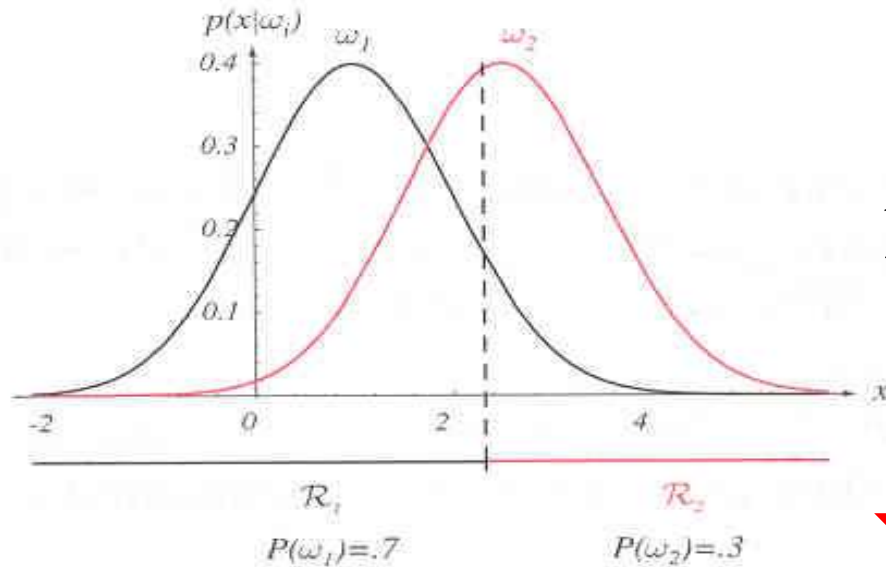


Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (3)

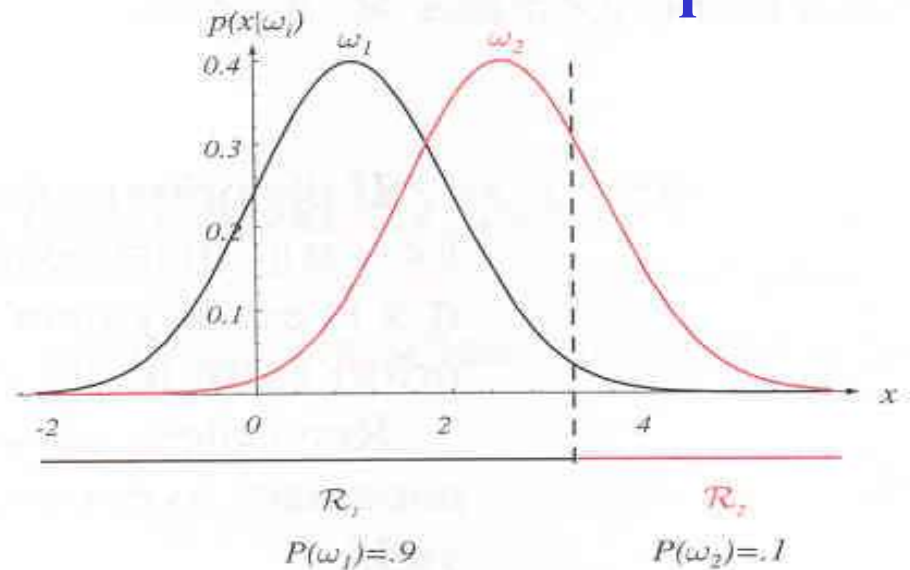
- Le funzioni discriminanti lineari definiscono un iperpiano passante per \mathbf{x}_0 ed ortogonale a \mathbf{w} :
dato che $\mathbf{w} = \boldsymbol{\mu}_i - \boldsymbol{\mu}_j$, l'iperpiano che separa R_i da R_j è *ortogonale* alla linea che unisce le medie.
- Dalla formula precedente si nota che, a parità di varianza, il prior maggiore determina la classificazione.



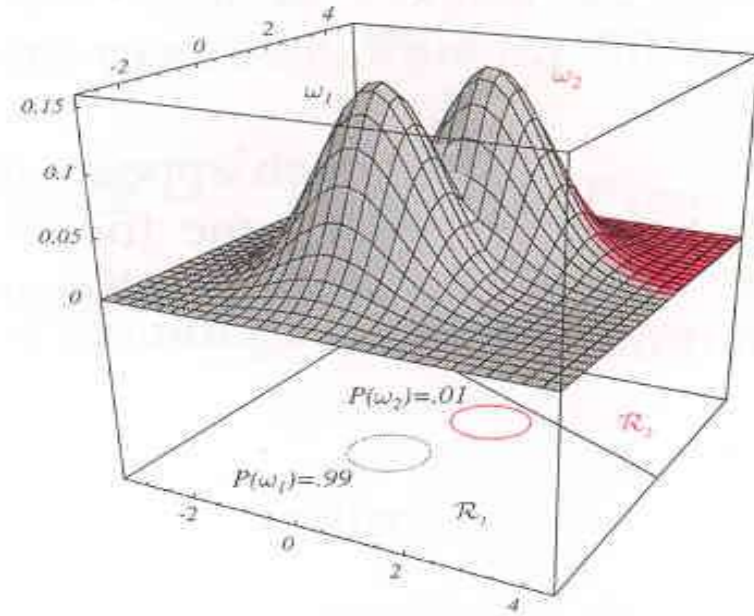
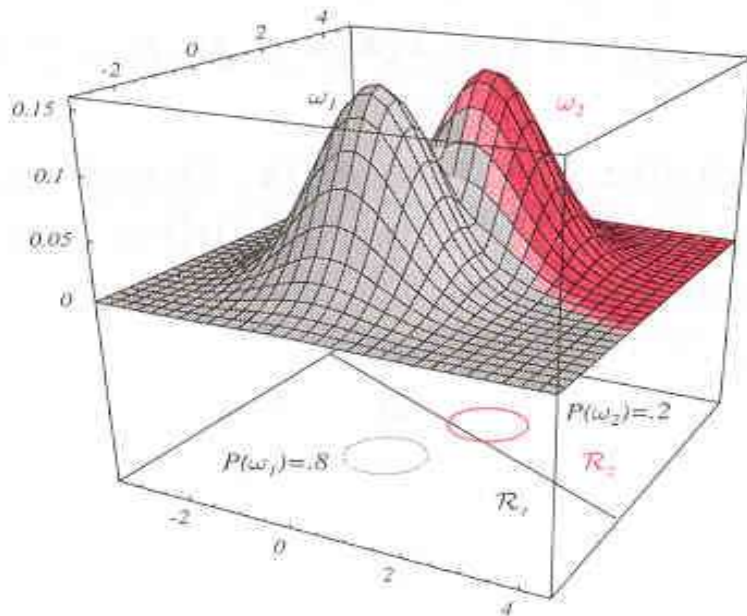
Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 I$ (4)



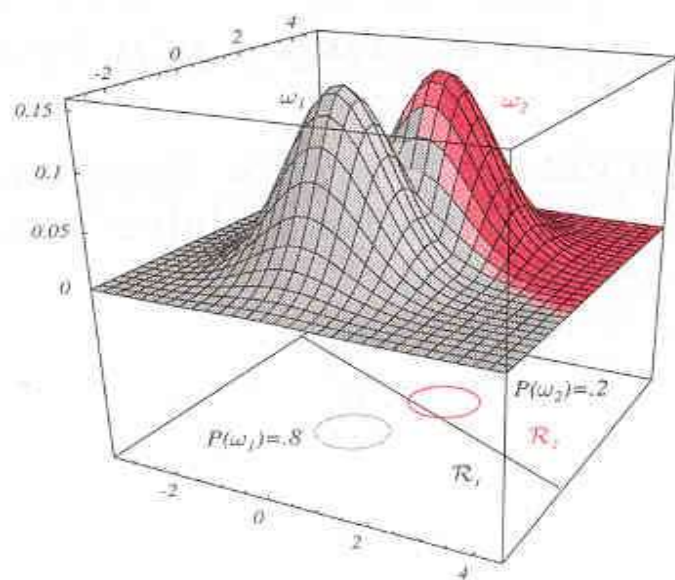
1-D



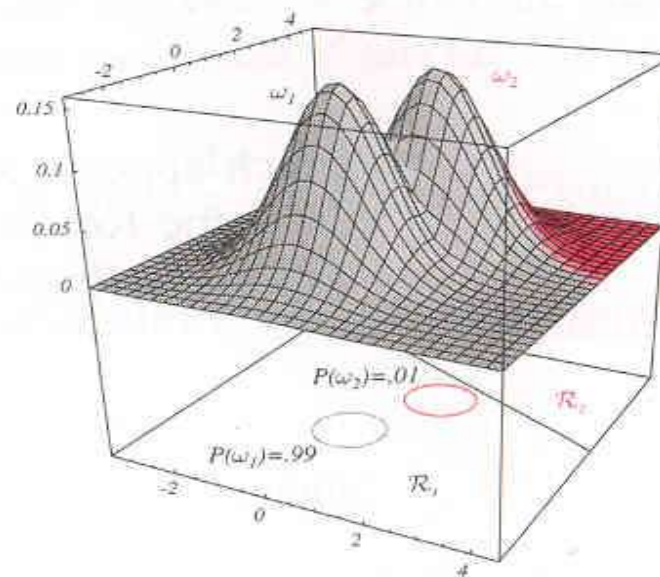
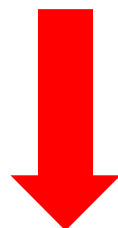
2-D



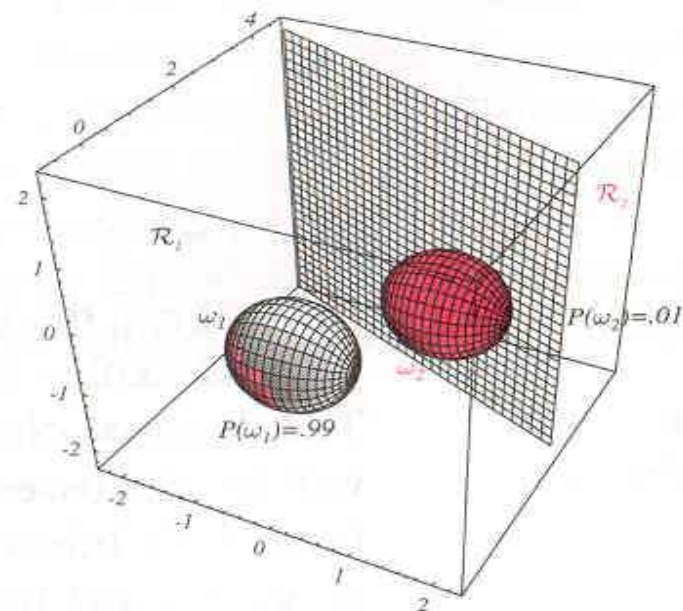
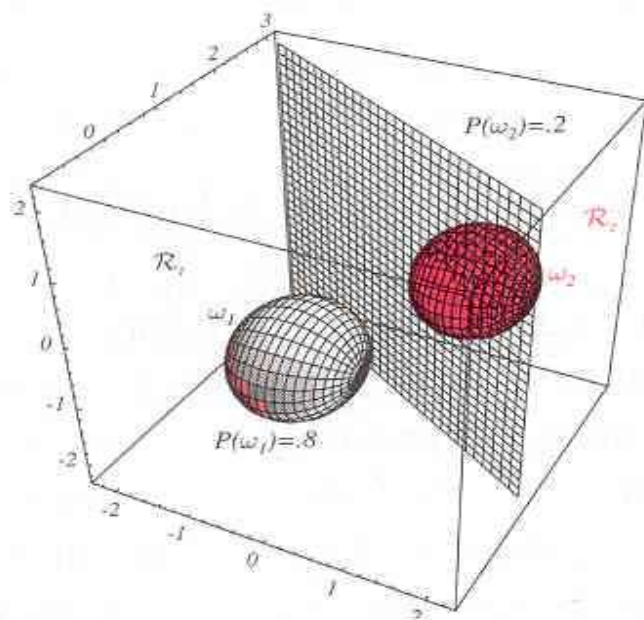
Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 I$ (5)



2-D



3-D



Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (6)

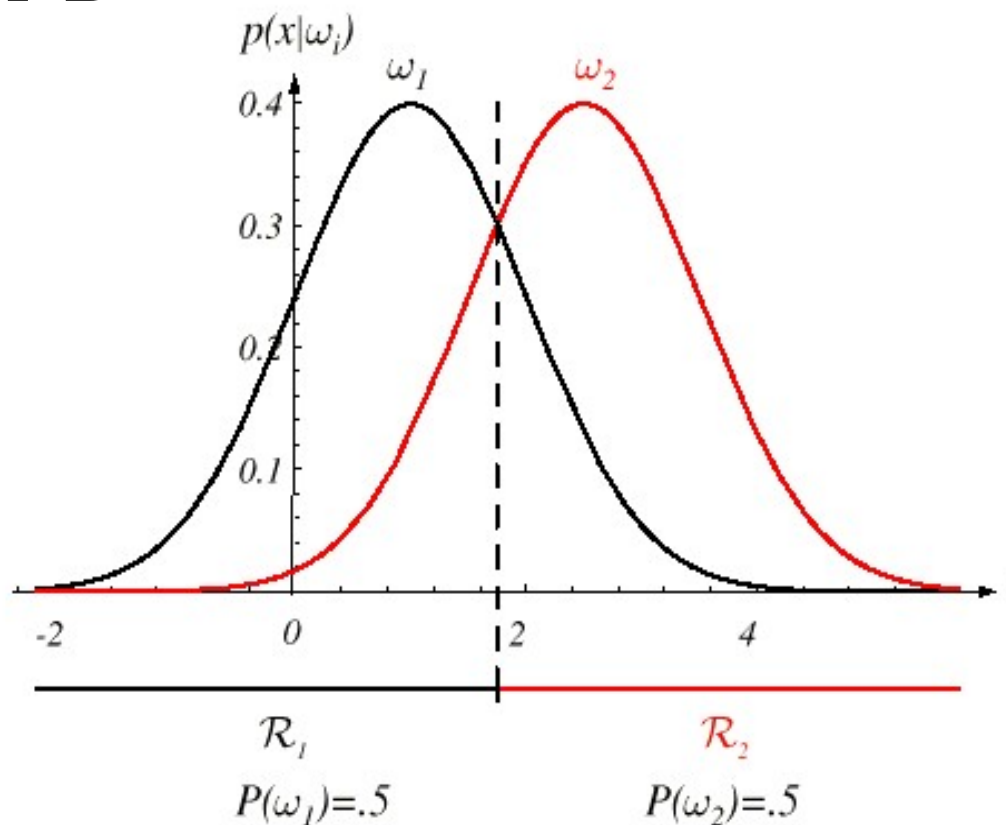
$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\sigma^2}{\|\boldsymbol{\mu}_i - \boldsymbol{\mu}_j\|^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$



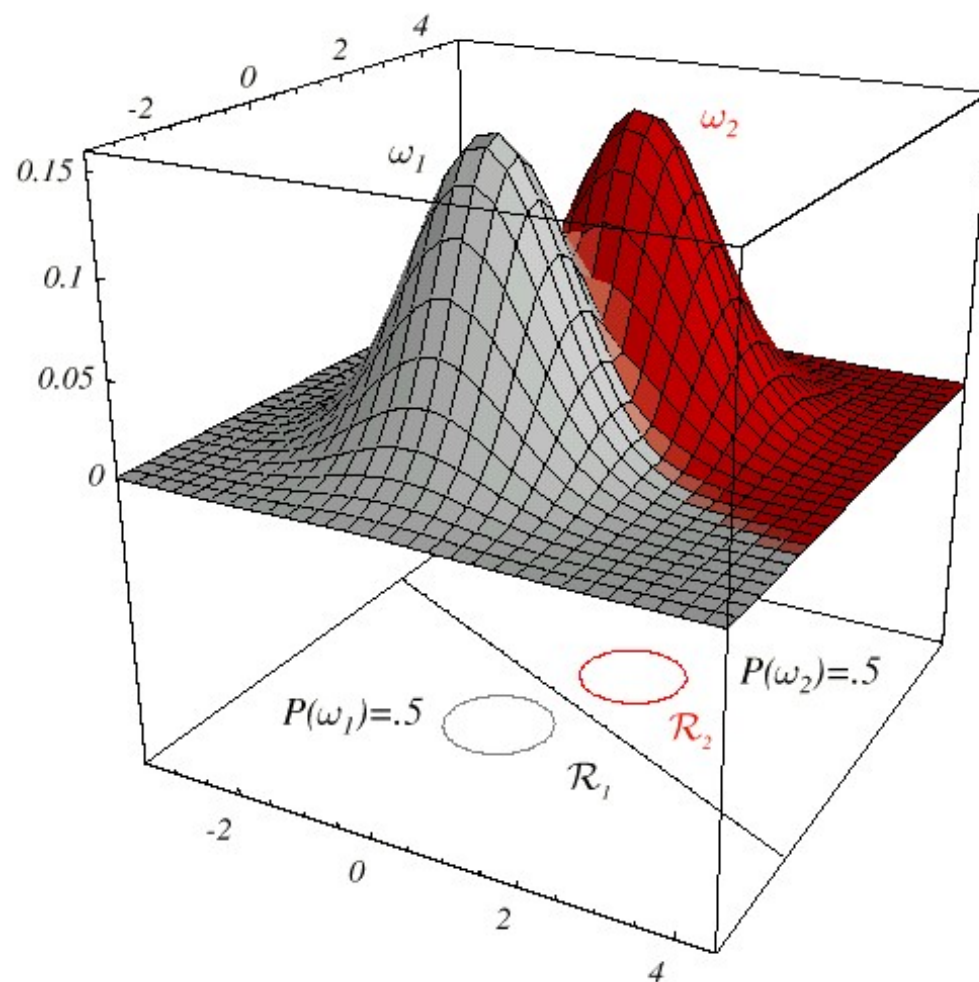
- NB.: Se le probabilità prior $P(\omega_i)$, $i=1, \dots, c$ sono *uguali*, allora il termine logaritmico può essere ignorato, riducendo il classificatore ad un ***classificatore di minima distanza***.
- In pratica, la regola di decisione ottima ha una semplice interpretazione geometrica
 - Assegna \mathbf{x} alla classe la cui media $\boldsymbol{\mu}$ è più vicina

Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (7)

1-D

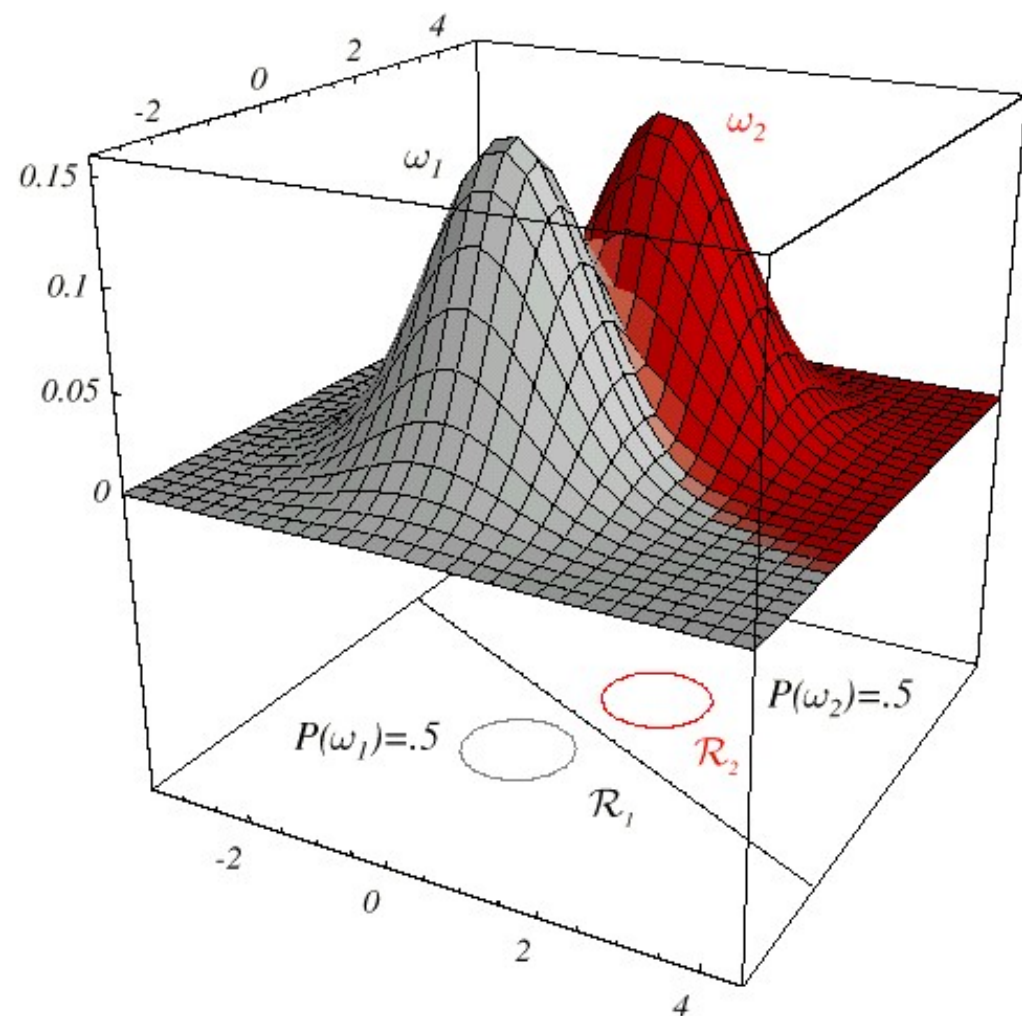


2-D

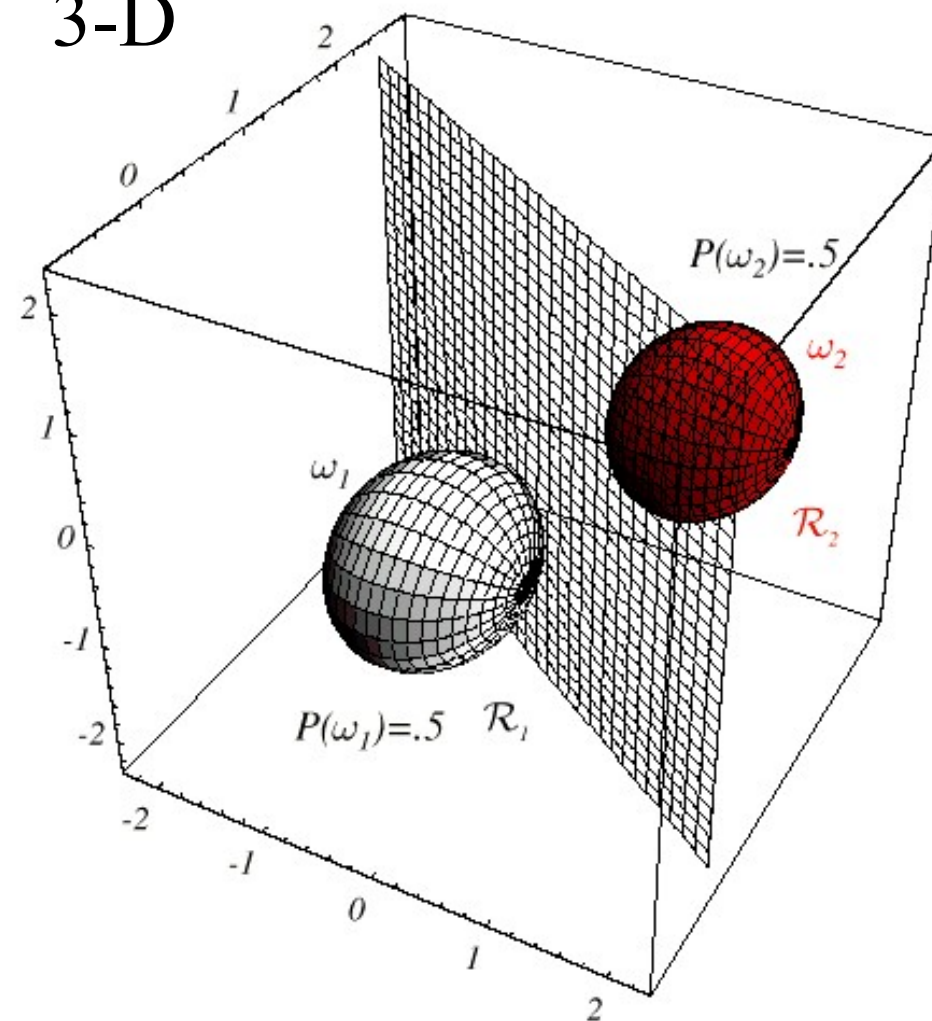


Funzioni discriminanti - Densità Normale $\Sigma_i = \sigma^2 \mathbf{I}$ (8)

2-D



3-D



Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$

- Un altro semplice caso occorre quando le matrici di covarianza per tutte le classi sono uguali, ma arbitrarie.
- In questo caso l'ordinaria formula

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma_i^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i)$$

può essere semplificata con

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^t \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) + \ln P(\omega_i)$$

che è ulteriormente trattabile, con un procedimento analogo al caso precedente (sviluppando il prodotto ed eliminando il termine $\mathbf{x}^t \Sigma^{-1} \mathbf{x}$)

Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (2)

- Otteniamo così funzioni discriminanti ancora lineari, nella forma:

$$g_i(\mathbf{x}) = \mathbf{w}_i^t \mathbf{x} + w_{i0}$$

dove

$$\mathbf{w}_i = \Sigma^{-1} \boldsymbol{\mu}_i$$

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma^{-1} \boldsymbol{\mu}_i + \ln P(\omega_i)$$

- Poiché i discriminanti sono lineari, i *confini di decisione* sono ancora iperpiani

Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (3)

- Se le regioni di decisione R_i ed R_j sono contigue, il confine tra esse diventa:

$$\mathbf{w}'(\mathbf{x} - \mathbf{x}_0) = 0,$$

where

$$\mathbf{w} = \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$$

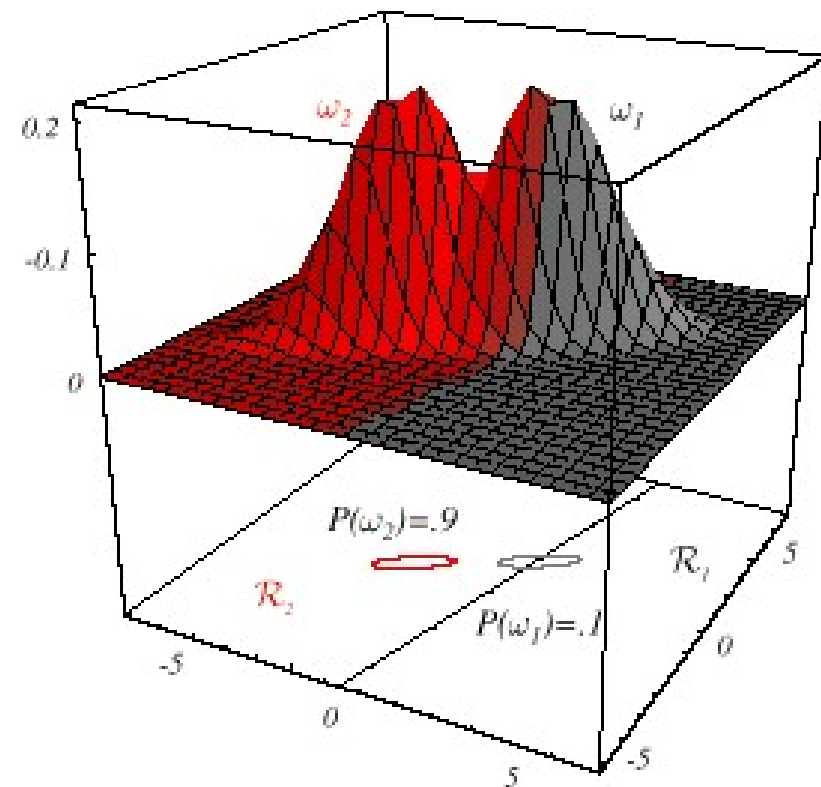
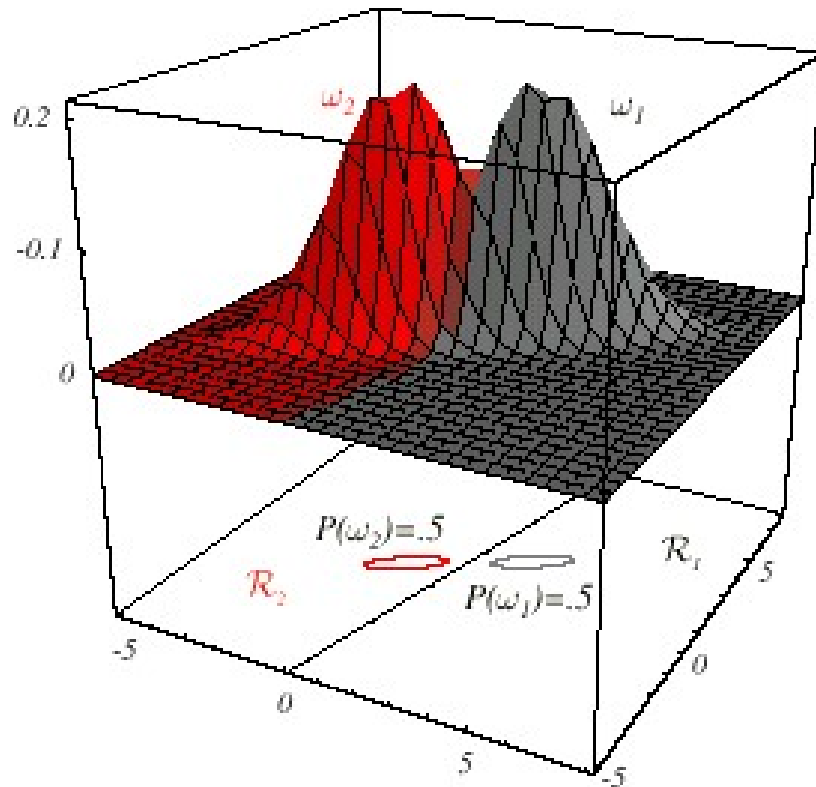
and

$$\mathbf{x}_0 = \frac{1}{2}(\boldsymbol{\mu}_i + \boldsymbol{\mu}_j) - \frac{\ln[P(\omega_i)/P(\omega_j)]}{(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)' \Sigma^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j).$$

Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (4)

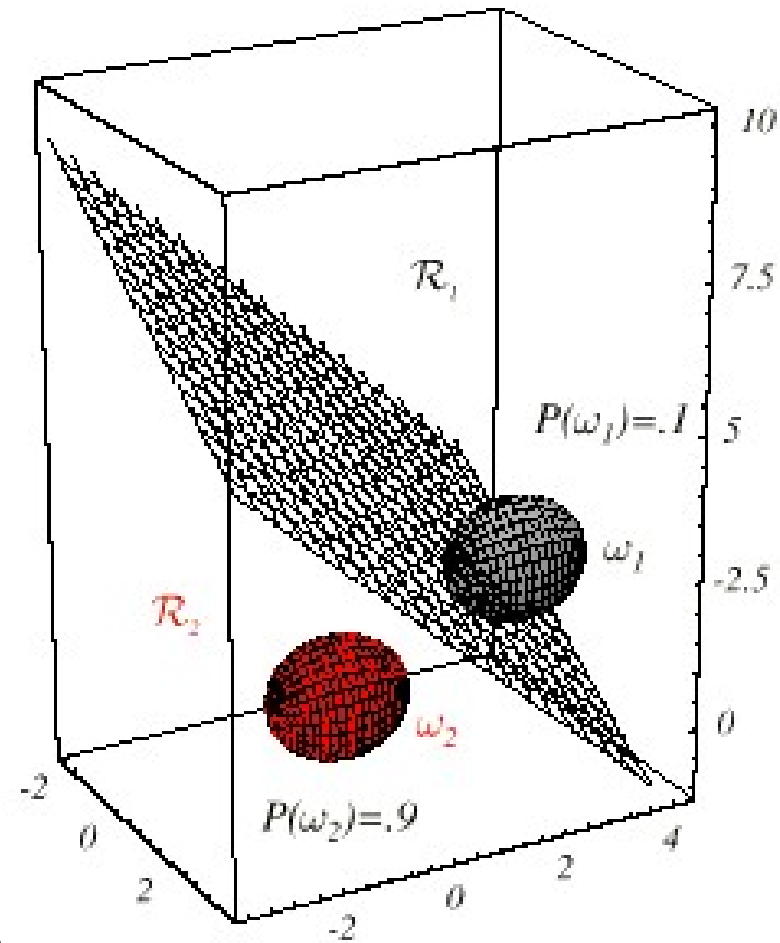
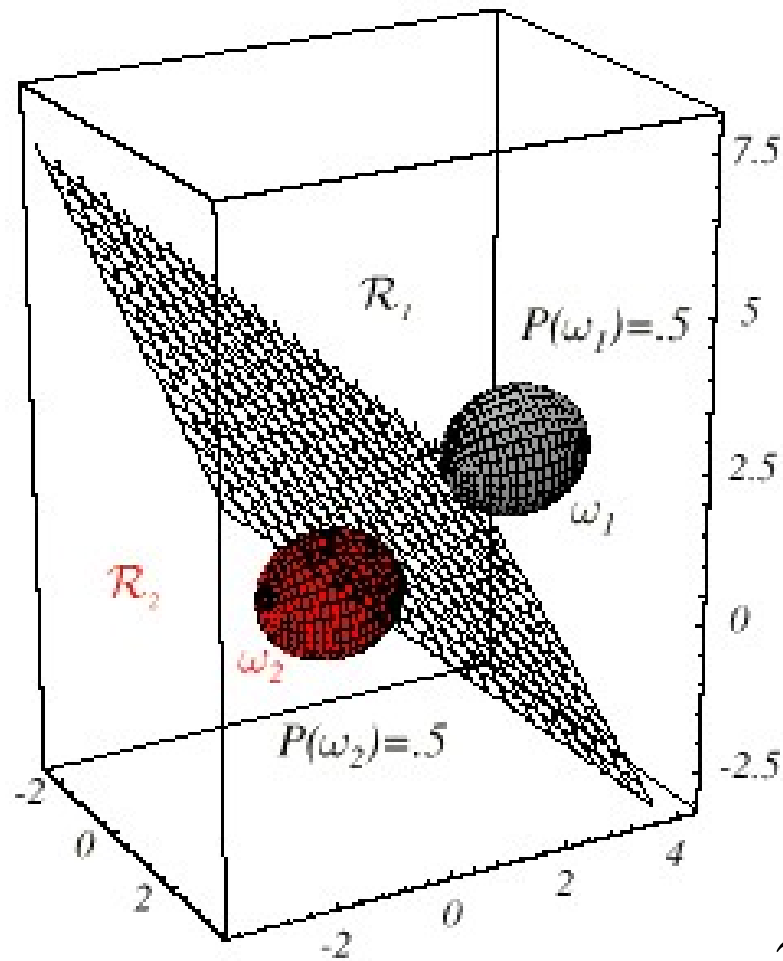
- Poiché \mathbf{w} in generale (differentemente da prima) non è il vettore che unisce le 2 medie ($\mathbf{w} = \mu_i - \mu_j$), l'iperpiano che divide R_i da R_j non è quindi ortogonale alla linea tra le medie; comunque, esso interseca questa linea in \mathbf{x}_0
- Se i *prior* sono uguali, allora \mathbf{x}_0 si trova in mezzo alle medie, altrimenti l'iperpiano ottimale di separazione si troverà spostato verso la media meno probabile.

Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (5)



2-D

Funzioni discriminanti - Densità Normale $\Sigma_i = \Sigma$ (6)



3-D

Funzioni discriminanti - Densità Normale Σ_i arbitraria

- Le matrici di covarianza sono differenti per ogni categoria;
- Le funzioni discriminanti sono inerentemente quadratiche;

$$g_i(\mathbf{x}) = \mathbf{x}^t \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^t \mathbf{x} + w_{i0},$$

where

$$\mathbf{W}_i = -\frac{1}{2} \Sigma_i^{-1},$$

$$\mathbf{w}_i = \Sigma_i^{-1} \boldsymbol{\mu}_i$$

and

$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^t \Sigma_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(\omega_i).$$

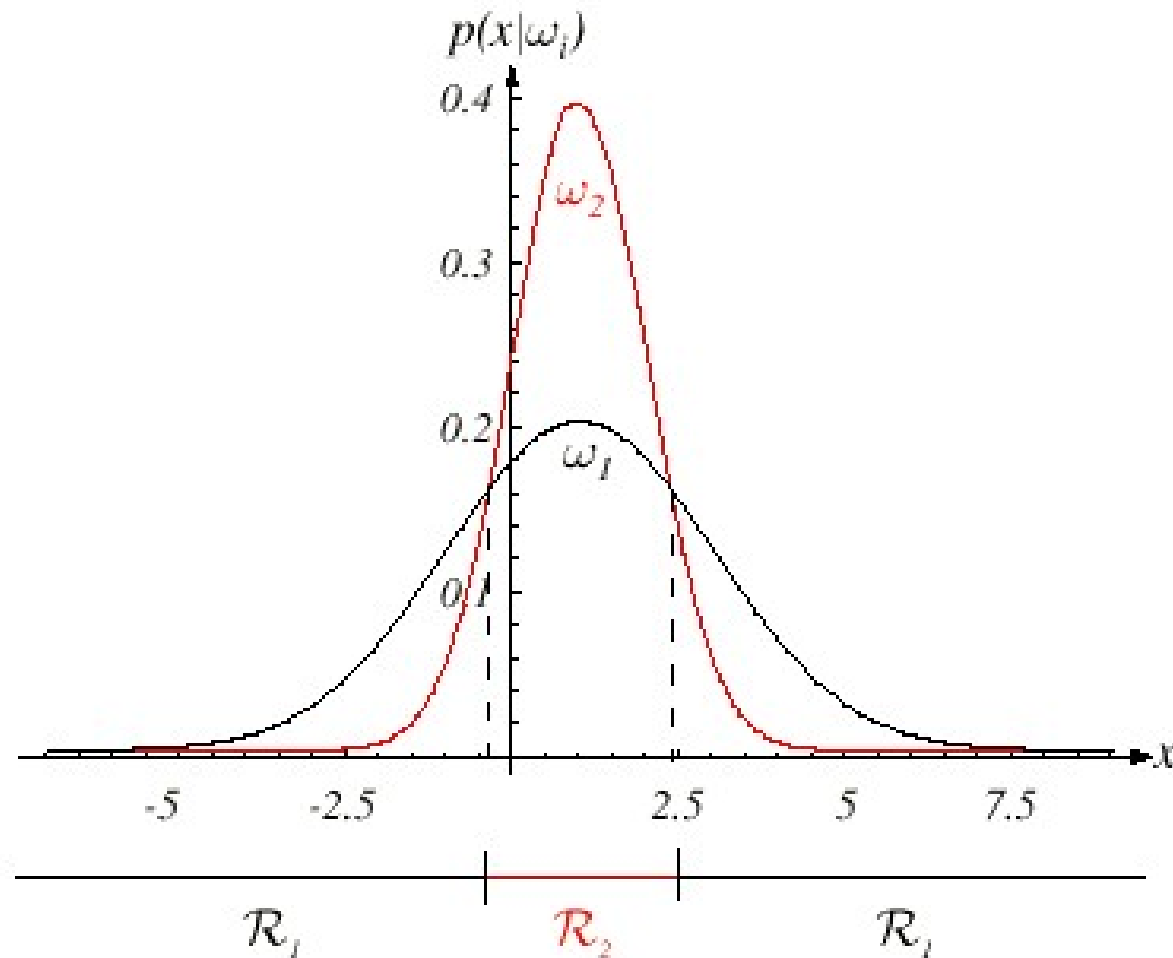
Funzioni discriminanti

Densità Normale Σ_i arbitraria (2)

- Nel caso 2-D le superfici di decisione sono *iperquadriche*:
 - Iperpiani
 - Coppia di iperpiani
 - Ipersfere
 - Iperparaboloidi
 - Iperiperboloidi di vario tipo
- Anche nel caso 1-D, per la varianza arbitraria, le regioni di decisione di solito sono non connesse.

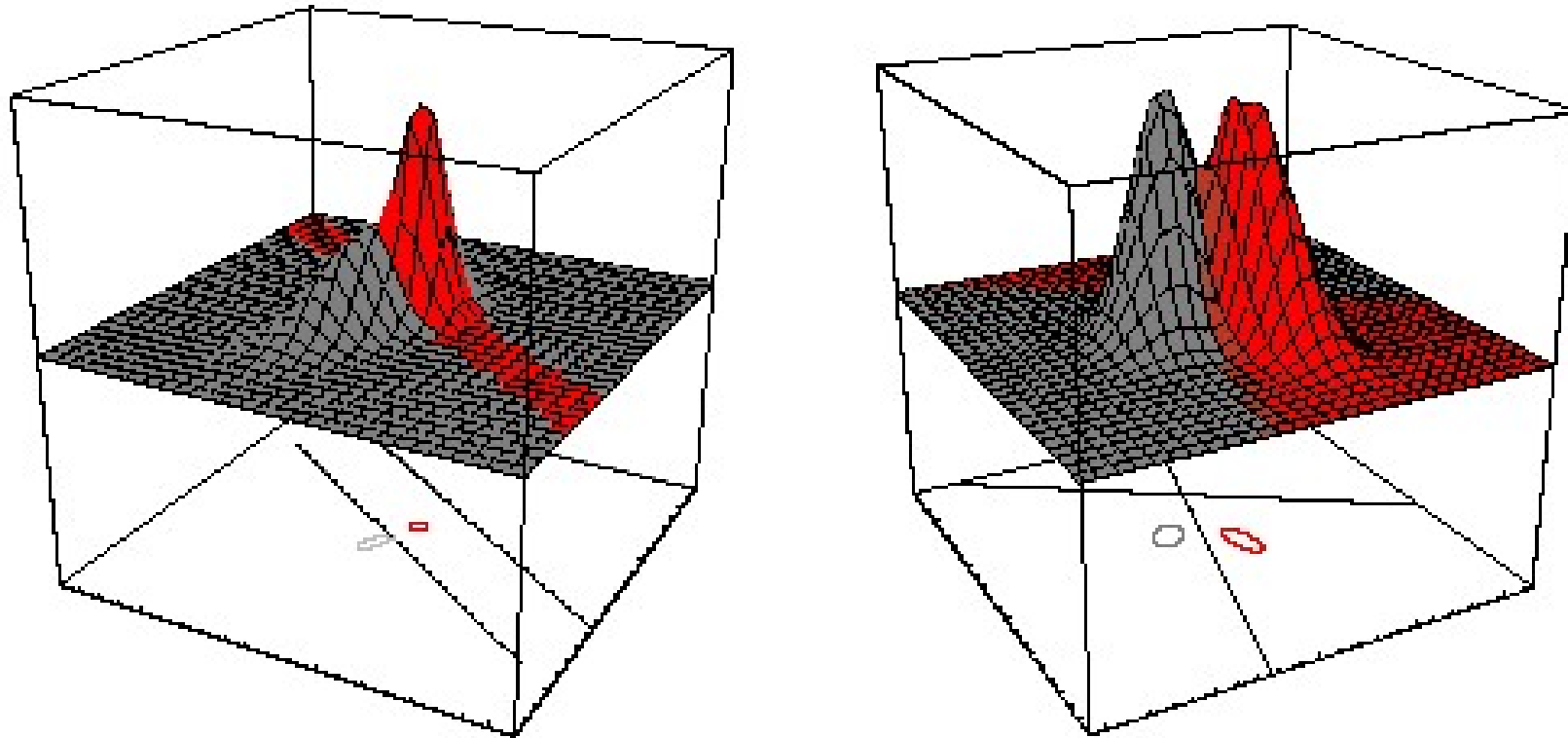
Funzioni discriminanti

Densità Normale Σ_i arbitraria (3)



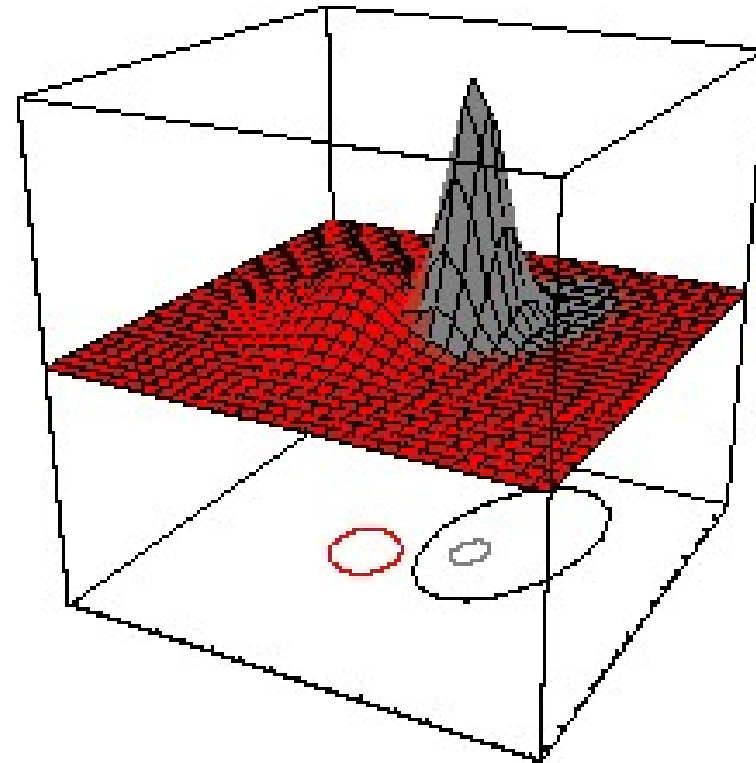
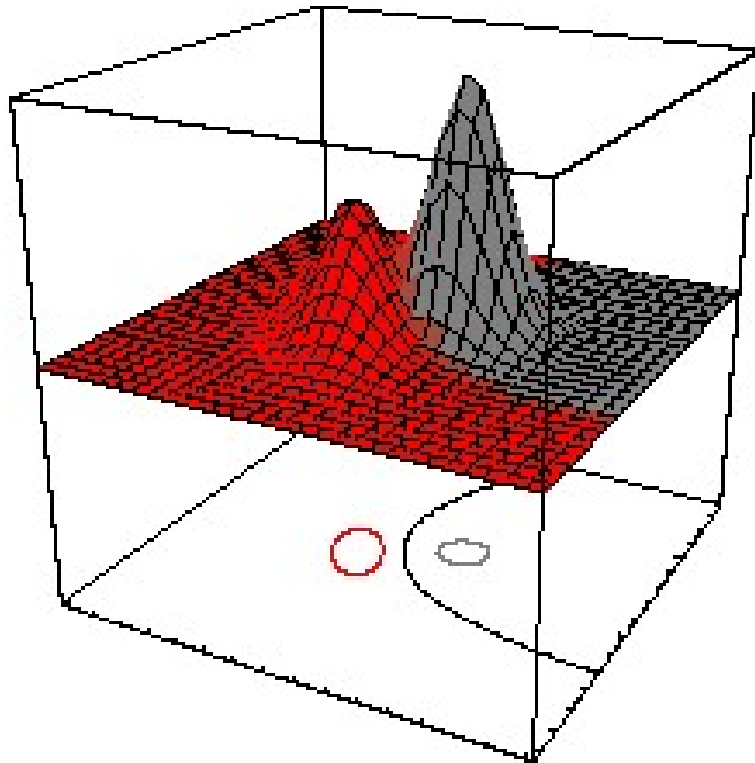
Funzioni discriminanti

Densità Normale Σ_i arbitraria (4)



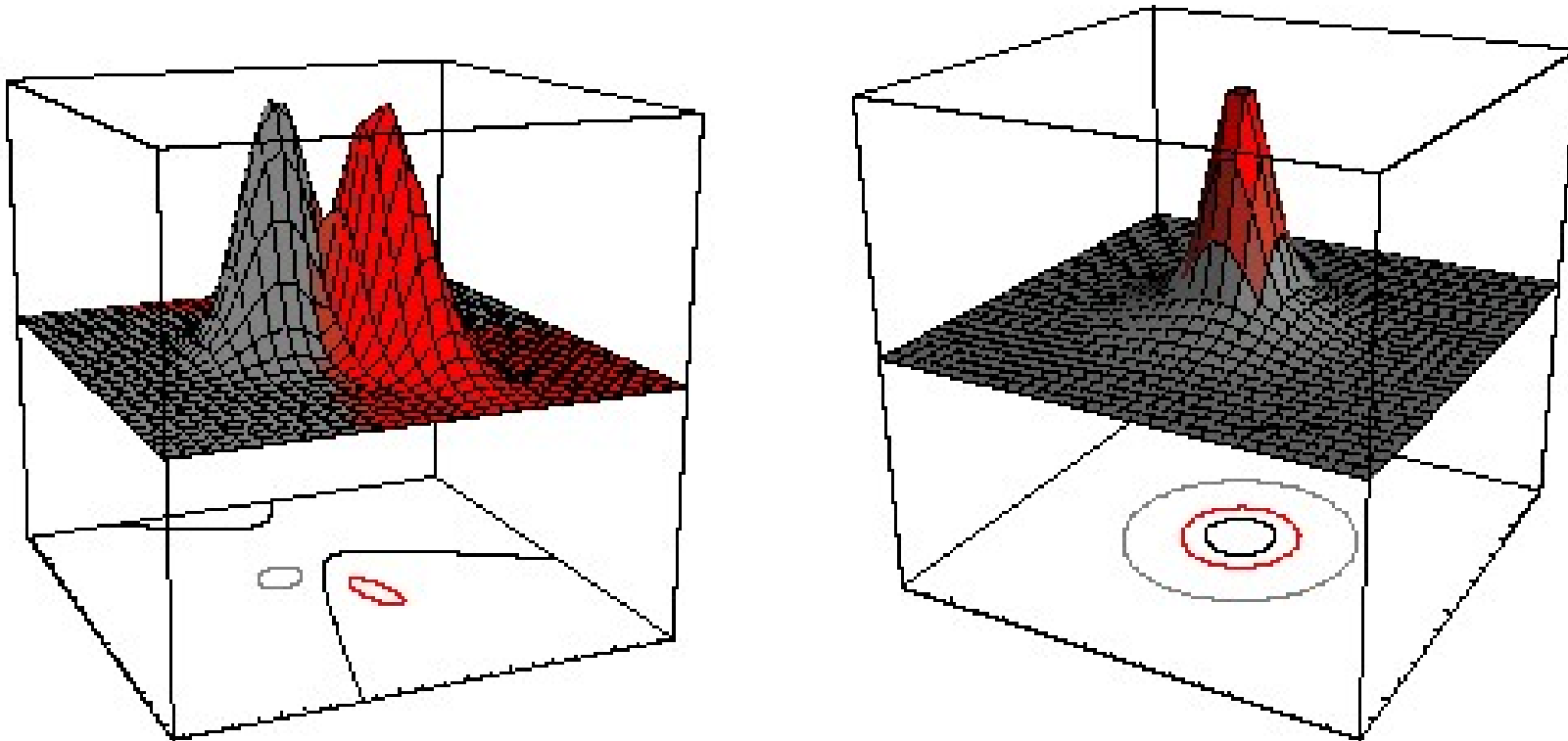
Funzioni discriminanti

Densità Normale Σ_i arbitraria (5)



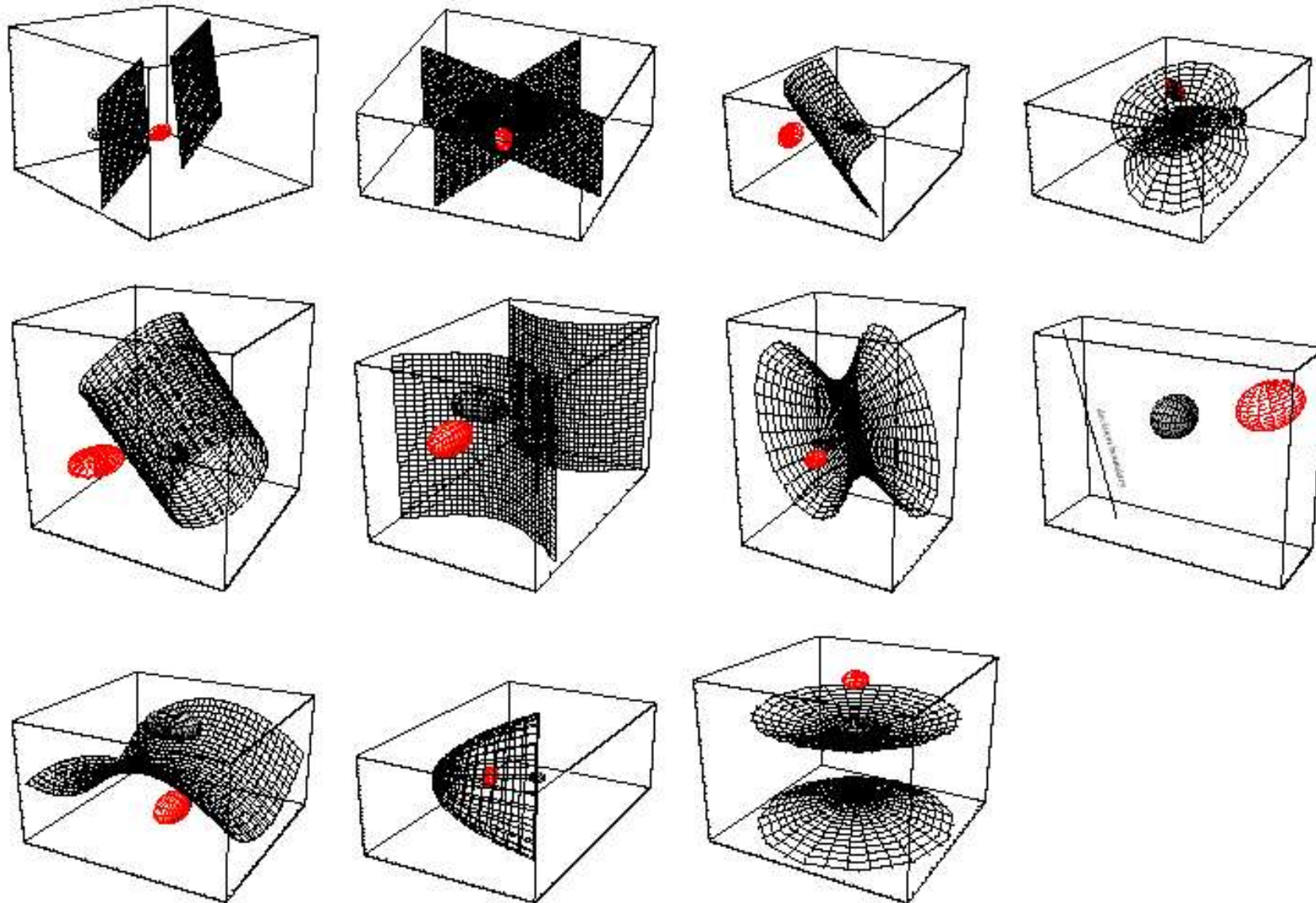
Funzioni discriminanti

Densità Normale Σ_i arbitraria (6)



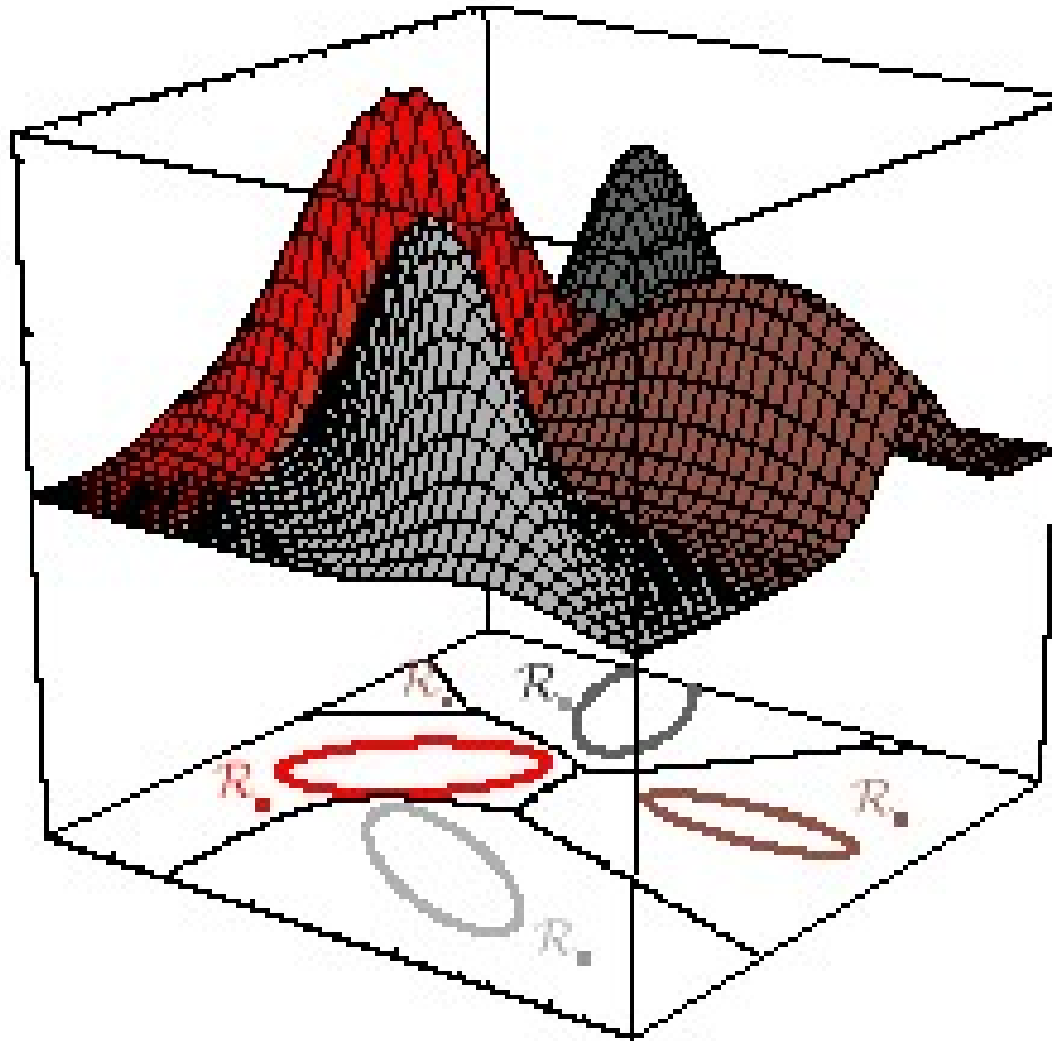
Funzioni discriminanti

Densità Normale Σ_i arbitraria (7)



Funzioni discriminanti

Densità Normale Σ_i arbitraria (8)



Riferimenti

- Libro Duda, fino a Sez. 2.6 compresa
- *No 2.3.1, 2.3.2*